Alignment of an Alternate Assessment with State Academic Standards:

Evidence for the Content Validity of the Wisconsin Alternate Assessment

Andrew T. Roach, *University of Wisconsin–Madison*
Stephen N. Elliott, *Peabody College of Vanderbilt University*
and
Norman L. Webb, *University of Wisconsin–Madison*

In this article, the authors describe an alignment and content analysis of the Wisconsin Alternate Assessment (WAA) for students with disabilities. The WAA is an assessment of the academic performance of students with significant disabilities and is an alternative to the traditional on-demand achievement test. Alternate assessments like the WAA are required by federal law and are expected to be aligned with state content standards. The primary purpose of this investigation was to determine the extent to which the WAA adequately measured the concepts and skill areas represented in Wisconsin's Model Academic Standards. The ratings of an expert panel ($N = 10$) that participated in the WAA Alignment Institute provided data regarding the alignment of WAA items to the standards. The expert panel's responses indicated that the WAA generally meets the multifaceted criteria developed by Webb (1997) and advocated by Title I reviewers for acceptable alignment between assessments and curriculum expectations as articulated in academic content standards.

For many students with disabilities, participation in state and district assessments involves taking existing standardized tests with testing accommodations. Some students (perhaps 0.5% to 2% of the student population), however, have disabilities that make their participation in state- and district-wide tests impractical and render the tests an inaccurate measure of their academic achievements. For example, a student with a developmental disability may not be able to understand and respond to items on a state's large-scale multiple-choice test. For such cases, the Individuals with Disabilities Education Act of 1997 (IDEA) required states to create and implement alternate assessment systems by July 1, 2000, and include the performance of students participating in alternate assessments in public accountability reporting.

The mandate to create alternate assessments has led states to propose a variety of methods for assessing students with significant disabilities. According to a survey of state special education directors conducted by Thompson and Thurlow (2003), the most common element in alternate assessments is a portfolio or body of evidence (23 of 50 states), followed by a rating scale or checklist (15 states), performance tasks or events (9 states), and Individualized Education Program (IEP) analysis (4 states). As the survey responses indicated, states are using multiple data collection methods to increase the validity of their alternate assessment systems. Moreover, many states' alternate assessment systems are in flux as modifications are made to respond to Title I reviews, No Child Left Behind Act (2001) legislation on adequate yearly progress (AYP), and demands to improve the reliability and validity of inferences based on alternate assessment results.

An Element of Inclusive Assessment Systems

Alternate assessments are an important component of each state's assessment system and, as such, are required to meet the federal requirements outlined in the Elementary and Secondary Education Act (2002). Specifically, the act, as amended by the No Child Left Behind Act of 2001, mandates that state assessments "be aligned with the State's challenging content and student academic performance standards, and provide coherent information about student attainment of such standards" (Elementary and Secondary Education Act, 2002). Many states have struggled to meet these requirements because the skills and concepts in the state academic standards

Address: Andrew T. Roach, Department of Educational Psychology, 1025 W. Johnson St., Madison, WI 53706; e-mail: atroach@wisc.edu

were deemed inappropriate or irrelevant for students with significant disabilities and the development of the alternate assessment was considered a special education function, precluding the involvement of general education curriculum and measurement experts.

The alignment between an assessment and the content it is meant to assess is an important piece of evidence in any validity argument. Lane (1999) outlined procedures for evaluating the validity of assessments designed to measure students' mastery of state academic standards. According to Lane, two forms of evidence are pertinent to determining the validity of these assessments: the extent to which the state assessment reflects the state's academic standards and the extent to which the curriculum offered to students reflects the academic standards. The purpose of this investigation was to provide evidence of the alignment between the Wisconsin Alternate Assessment (WAA) for students with disabilities and Wisconsin's Model Academic Standards. By establishing the alignment and curricular relevance of the WAA, this investigation provided evidence of the validity of the WAA results as a measure of students' mastery of the academic concepts and skills outlined in the Wisconsin Model Academic Standards. In addition, the investigation demonstrated the use of a formal procedure to establish the alignment of an alternate assessment.

Enhancing Alternate Assessment in Wisconsin

In Wisconsin, the original alternate assessment involved a review of student performance similar to what might typically be part of a reevaluation procedure or an IEP process. The Wisconsin Department of Public Instruction stated that the alternate assessment could consist of any of the following elements: school records; the most recent evaluation data; formal and informal assessments conducted by team members; reports by parents, general education teachers, and special education teachers; classroom work samples; and other information available to the IEP team (Elliott, 2001). In addition, for the IEP review process to be considered an alternate assessment, it had to be (a) a comprehensive, recent, and representative review of student performance; (b) conducted in the same general time frame as statewide large-scale testing; and (c) aligned with the state's general education standards.

Although this approach to alternate assessment appeared to meet the IDEA guidelines for the participation of students with disabilities in assessment, some educators and policymakers identified concerns with having students involved in primarily idiographic assessments. According to Thurlow et al. (1996), the problems with this approach are twofold: individual students' attainment of IEP goals are not easily aggregated to determine systemwide accountability, and IEP goals and objectives should not represent the total curriculum for a student. Moreover, because functional and adaptive behaviors are often the focus of IEP goals for students with significant disabilities, many alternate assessments under the original approach would not have reflected the range of knowledge and skills identified by Wisconsin's Model Academic Standards.

In response to these concerns and to the questions raised by a Title I review completed by the U.S. Department of Education in February 2001, Wisconsin began the process of designing and implementing an enhanced alternate assessment that would provide more structure to teachers, clearer alignment to the state's academic standards, and more manageable data on students' performance. This enhanced version of the WAA includes a behavior rating scale based on the state's alternate performance indicators (APIs), a downward extension of the state's academic standards. In addition, the WAA includes an overall scoring continuum for each core subject area (i.e., reading, language arts, math, social studies, and science), which allows student performance to be categorized in a manner similar to the proficiency levels used to describe students' performance on the Wisconsin Knowledge and Concepts Examinations (WKCE) at Grades 4, 8, and 10.

Measures of Access to the General Curriculum

IDEA clearly mandates that students with disabilities have access to the general education curriculum and academic standards. Specifically, one of the final regulations under IDEA (34 C.F.R. § 300.347) requires that students' IEPs consider how the students will access the general education curriculum. This regulation further requires that all students participate in statewide and districtwide assessments and all students have opportunities and instruction that allow them to make progress toward state and district academic standards.

This emphasis on attaining academic achievement represents a dramatic departure from the curriculum and inclusion practices that traditionally have been implemented with many students with significant disabilities. Early considerations of mainstreaming and least restrictive environment (LRE) often focused on the socialization and self-esteem benefits for students with significant disabilities. More recent practices have maintained the focus on relationships and self-concept while adding an emphasis on exposure to the general curriculum and the broader school experience (Ford, Davern, & Schnorr, 2001). IDEA, however, demands even greater access to the general education curriculum. Students must have instruction and accommodations that promote their progress, no matter how modest, toward the education expectations of the larger student population.

If the alternate assessments are intended to function as one element of a larger accountability system and to measure progress toward the same education expectations applied to the larger student population, then a state's general education academic standards should form the foundation for the alternate assessment. IDEA seems to provide support for the design of alternate assessments as an extension or modification of

states' standards-based assessment systems. Although the primary purpose of state and district assessments is to measure students' progress toward major, agreed-on academic expectations, in many cases important instructional experiences and opportunities are not reflected in these assessments (Ford et al., 2001). Indeed, much as we do not expect multiple-choice standardized tests to measure the entire scope of curriculum and instruction provided to general education students, we should not expect alternate assessments to reflect every element of the school experience of students with significant disabilities.

Alignment Among Standards, Assessments, and Classroom Practices

Effective schooling is based on the coordination of three components of the educational environment: curriculum, instruction, and assessment (Elliott, Braden, & White, 2001; Webb, 1997, 2002; Webb, Horton, & O'Neal, 2002). The degree to which these elements work together toward student learning is *alignment*—and the foundation of standards-based education reform. Alignment is the extent "to which expectations and assessments are in agreement and serve in conjunction with one another to guide the system toward students learning what they are expected to know and do" (Webb, 2002, p. 1). The development and implementation of large-scale assessment programs represent one approach to aligning classroom instruction with state curriculum standards.

The Council of Chief State School Officers (CCSSO) has identified four preferred models as frameworks for states planning and conducting alignment studies: (a) the Webb model, (b) the Surveys of Enacted Curriculum (SEC) model, (c) the "Achieve" model, and (d) the Council for Basic Education (CBE) model (CCSSO, 2002). This investigation used Webb's alignment model, which provides a series of statistics that indicate the match between the content in the state's academic standards and the content covered by the state assessment. Because the Webb model was used in a previous study to determine the alignment of Wisconsin's large-scale assessment (the WCKE) with the state's academic standards, the application of this model to the WAA was intended to provide policymakers with comparative data on the alignment of two elements of the state's assessment system. Beyond its application in Wisconsin, the Webb model has been used to judge the alignment between standards and assessments for language arts, mathematics, social studies, and science in more than 10 states. These states have used information from the Webb alignment analyses to modify assessments, to alter standards, and to verify the extent to which these documents are directed toward common expectations for learning.

Webb (2002) and Webb, Horton, and O'Neal (2002) represent two applications of Webb's method for analyzing the alignment of assessments and curriculum standards. In these examples, panels of curriculum experts were trained to use an analytic process and heuristics to rate the alignment between states' assessment systems and academic standards. Analyses of the panel members' responses provided information on the assessments' attainment of the following alignment criteria: (a) categorical concurrence, (b) balance of representation, (c) range-of-knowledge correspondence, and (d) depth-of-knowledge consistency. The first three criteria measure the correspondence between skills and concepts covered by the state's content standards and objectives (i.e., performance standards) and the skills and concepts tested by an assessment. Categorical concurrence indicates whether the same or consistent categories of content appear in both the content standards and the assessment items. Range-of-knowledge correspondence indicates whether the span of knowledge expected of students by a standard is the same as, or corresponds to, the span of knowledge that students need to answer correctly assessment items or activities. Balance of representation provides an index of the degree to which one curriculum objective is given more emphasis on the assessment than another. Conversely, depth-of-knowledge consistency is intended to represent the level of complexity required by the objectives and assessment items. The depth-of-knowledge criterion indicates whether what is elicited from students on an assessment is as complex for the content area as what students are expected to know and do as stated in the model academic standards. Figure 1 illustrates how Webb's criteria measure the relationships among a state's academic standards and large-scale assessment system.

By focusing on the alignment of the WAA with Wisconsin's academic standards, this investigation provided content-related evidence for the validity of the WAA. Specifically, the investigation sought to answer the following question: Does the WAA adequately measure the concepts and skill areas represented in Wisconsin's Model Academic Standards? The ratings of the expert panel that participated in the WAA Alignment Institute provided information about the correspondence between WAA items and the Wisconsin Model Academic Standards on multiple criteria. Because Wisconsin's APIs served as the framework for WAA item development, analysis of the alignment panel's ratings was expected to indicate that each WAA subject domain scale met the criteria identified by Webb for categorical concurrence, range of knowledge, and balance of representation. With regard to the panel's responses to the depth-of-knowledge rating, we expected a low overall rating for the WAA subject domain scales. The low overall depth-of-knowledge rating would represent a departure from previous alignment studies using expert panel ratings of general large-scale assessments (Webb, 2002; Webb et al., 2002). Although similar depth-of-knowledge ratings for curriculum objectives and assessment items are desirable, items on alternate assessments are generally intended to be less complex than items in the general education academic standards and on the corresponding large-scale assessment. In the case of the WAA

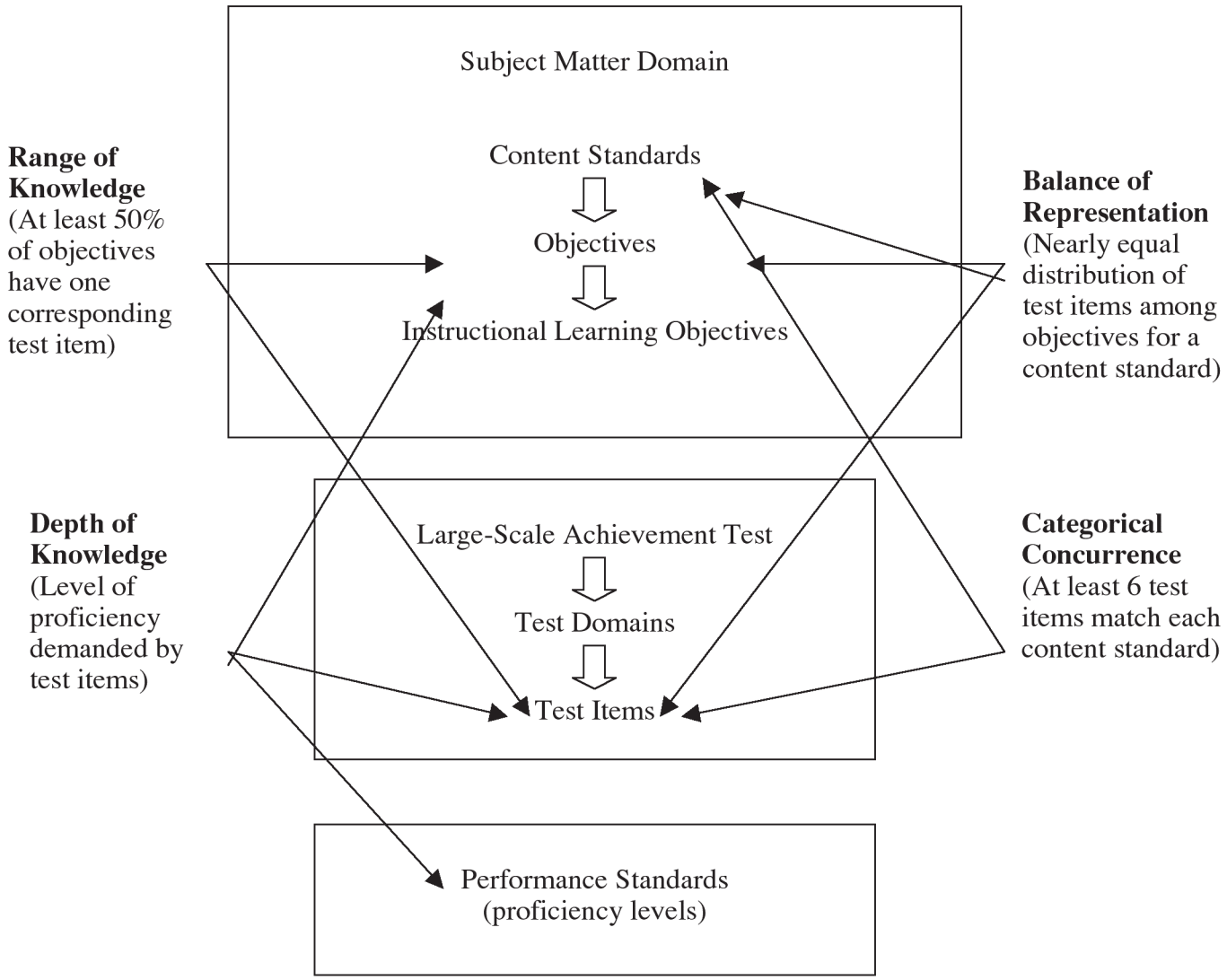

FIGURE 1. Illustration of how Webb's Alignment Criteria measure the correspondence between a state's academic standards and large-scale assessment systems.

rating scale, items were based on the state's Alternate Performance Indicators (APIs), which are downward extensions of the state's Model Academic Standards and intended to be at a level more indicative of the curricular and instructional needs of students with significant disabilities. Thus, although WAA items represent the range of concepts and skills outlined in the state academic standards, these items address skills thought to be less complex and more appropriate for students functioning developmentally several years below their chronological peers.

Given the stated purpose of the investigation and the research question, a method was needed to evaluate the alignment between WAA items and Wisconsin's Model Academic Standards.

Method

Information about the alignment of the WAA instrument with Wisconsin's Model Academic Standards for each of the subject domains assessed by the WCKE (the state's large-scale assessment) was collected during the WAA Alignment Institute conducted on June 13 and 14, 2002.

Participants

The alignment review panel ($N = 10$) consisted of special education teachers, personnel from DPI, and graduate students who participated in the 2-day WAA Alignment Institute conducted at the University of Wisconsin–Madison. Members of

the alignment panel had extensive understanding of testing and measurement, special education policy, and education accountability systems. The special education practitioners and state administrators who participated in the alignment panel were selected because of their experience with instruction and assessment for students with low-incidence disabilities. In addition, the graduate students who participated in the panel had completed a five-course assessment sequence that included training in the assessment of students with low-incidence disabilities and the participation of students with disabilities in state accountability systems. The majority of panel members also had participated in the WAA field trial (spring 2002), and all panel members were familiar with the instrument and its application prior to the Alignment Institute.

Instruments

The Wisconsin Alternate Assessment (WAA). The WAA is a part of the Wisconsin Student Assessment System and is designed to assess the academic performance of students with disabilities who cannot meaningfully participate in the general test (WKCE) even with accommodations. The WAA used in the spring 2002 field trial consisted of 128 Likert-scale items that required teachers to rate students' performance of a skill or understanding of a concept on a 4-point scale ranging from *nonexistent* (0) to *proficient/generalized* (3). In addition, teachers could rate items *not applicable* (NA) if they determined the item was not relevant to the student's educational needs. The WAA items were organized into five scales that assessed students' performance in each core academic subject: reading, language arts, math, social studies, and science.

To determine a student's overall performance level score for a subject domain, teachers were asked to review the results of their ratings for the domain and select the performance level descriptor from the domain developmental continuum. Each subject domain on the WAA has a 4-level, criterion-referenced, developmental continuum that characterizes performance of knowledge and skills along the path toward functioning at or near grade level in the regular curriculum. Thus, for each domain assessed, a student's performance was summarized as Prerequisite Skill Level 1 (Minimal), Prerequisite Level 2 (Basic), Prerequisite Level 3 (Proficient), or Prerequisite Level 4 (Advanced).

Table 1 presents a summary of statistics that describe the technical characteristics (e.g., central tendencies, score distributions, reliability indices) of each WAA content domain resulting from the spring 2002 WAA field trial investigation. These results should be interpreted with caution given that they are based on the ratings of a relatively small sample (40 students). The majority of students (83%) included in the field trial were identified by their teachers as having a cognitive

TABLE 1. Descriptive Statistics for WAA Subject Domain Scales

Descriptive & statistical indices	Reading	Language arts	Mathematics	Science	Social studies
Total No. of items/maximum possible score	23/69	26/78	29/87	21/63	29/87
Mean raw score	24.3	28.6	28.7	15.1	27.7
Median	20	24.0	23.5	14.5	26.0
Mode	2.0	2.0	3.0	3.0	4.0
SD	19.0	19.3	22.4	12.5	21.8
SEM	3.7	3.7	3.6	3.7	3.7
Scores associated with percentiles:					
25th percentile	6.0	12.0	7.8	3.0	7.0
50th percentile	20.0	24.0	23.5	14.5	26.0
75th percentile	44.0	49.0	49.0	23.0	43.0
Percentage of students at each performance level					
Prerequisite Skill 1	30.8	30.8	28.2	53.8	46.2
Prerequisite Skill 2	30.8	33.3	30.8	25.6	25.6
Prerequisite Skill 3	28.2	33.3	35.9	7.7	17.9
Prerequisite Skill 4	10.3	2.6	2.6	0	0
Coefficient alpha	0.98	0.97	0.98	0.96	0.98

Note. *SD* = standard deviation; *SEM* = standard error of measurement.

disability. The sample also included slightly more elementary school students (45%) than middle school (32.5%) or high school students (22.5%).

Procedure

The alignment coding process entailed panel members rating the alignment between the WAA rating scale items and Wisconsin's Model Academic Standards. The primary role of the panel members was to complete the following tasks:

1. Reach consensus on a depth-of-knowledge level rating for each objective in the Model Academic Standards.
2. Rate the depth-of-knowledge level of each item on the WAA rating scale.
3. Identify the one or two objectives from the Model Academic Standards to which each WAA item corresponds.

Before independently completing their ratings, panel members were trained to identify the depth-of-knowledge level for curriculum objectives (i.e., performance standards) and WAA items. This training included a review of the four general depth-of-knowledge levels outlined in Table 2. Specific descriptions for depth-of-knowledge levels for each of the subject domains covered by the WAA were developed, using examples from previous alignment analyses conducted on large-scale assessments as models (Webb, 2002; Webb et al., 2002). Panel members reached consensus on the depth-of-knowledge levels for curriculum objectives in each content domain before completing their individual ratings of WAA items in that content domain. Working as a group to reach consensus on the depth-of-knowledge levels for each objective provided an opportunity for discussion of the rating criteria, resulting in calibration of panel members' understanding of the depth-of-knowledge rating process (Webb, 2002).

Following this calibration process, panel members were asked to assign a depth-of-knowledge rating to each assessment item on a randomly ordered list of WAA items. Panel members' individual responses were recorded on a series of coding sheets, which provided columns for rating each WAA item on the depth-of-knowledge criteria and indicating the corresponding objectives for each item. WAA items were presented in random order instead of in the order in which they appear on the WAA rating scale, which corresponds with the organization of the state's Model Academic Standards. Figure 2 provides an example of one of the alignment coding sheets and an illustration of how panel members coded items.

If panel members had difficulty deciding between two levels for an objective or a WAA item (e.g., between a rating of 1 or 2), they were instructed to choose the higher of the two levels. After assigning the depth-of-knowledge rating for each item, panel members completed the coding sheets by identifying the one or two objectives that corresponded to the item.

The alignment coding process is not designed to produce exact agreement among members of the expert panel. In fact, variance in ratings may represent valid differences in opinion that reflect a lack of clarity in the objectives or the robustness of assessment items that could reasonably correspond to more than one curricular objective (Webb, 2002).

The alignment process completed by panel members consisted of depth-of-knowledge ratings for content standard objectives and WAA rating scale items and ratings of the cor-

TABLE 2. Depth-of-Knowledge Levels

Level	Description
Level 1: Recall	This level includes the recall of such information as a fact, definition, term, or simple procedure, as well as the ability to perform a simple algorithm or apply a formula.
Level 2: Skill/Concept	This level includes the engagement of some mental processing beyond a habitual response. A Level 2 assessment item requires students to make some decisions as to how to approach a problem or activity. Keywords that distinguish a Level 2 item or task include *classify, organize, estimate, make observations, collect and display data,* and *compare data.*
Level 3: Strategic thinking	This level includes items that require reasoning, planning, using evidence, and thinking at a higher level than the previous two levels. In most instances, requiring students to explain their thinking is a Level 3 attribute. Students might also be required to make conjectures or determine a solution to a problem with multiple correct answers.
Level 4: Extended thinking	This level includes items that require complex reasoning, planning, developing, and thinking, most likely during an extended time. At Level 4, the cognitive demands of the task should be high, and the work should be very complex. Students should be required to make connections both within and between subject domains. Level 4 activities include designing and conducting experiments; making connections between a finding and related concepts; combining and synthesizing ideas into new concepts; and critiquing literary pieces and experimental designs.

Note. Adapted from Webb (2002).

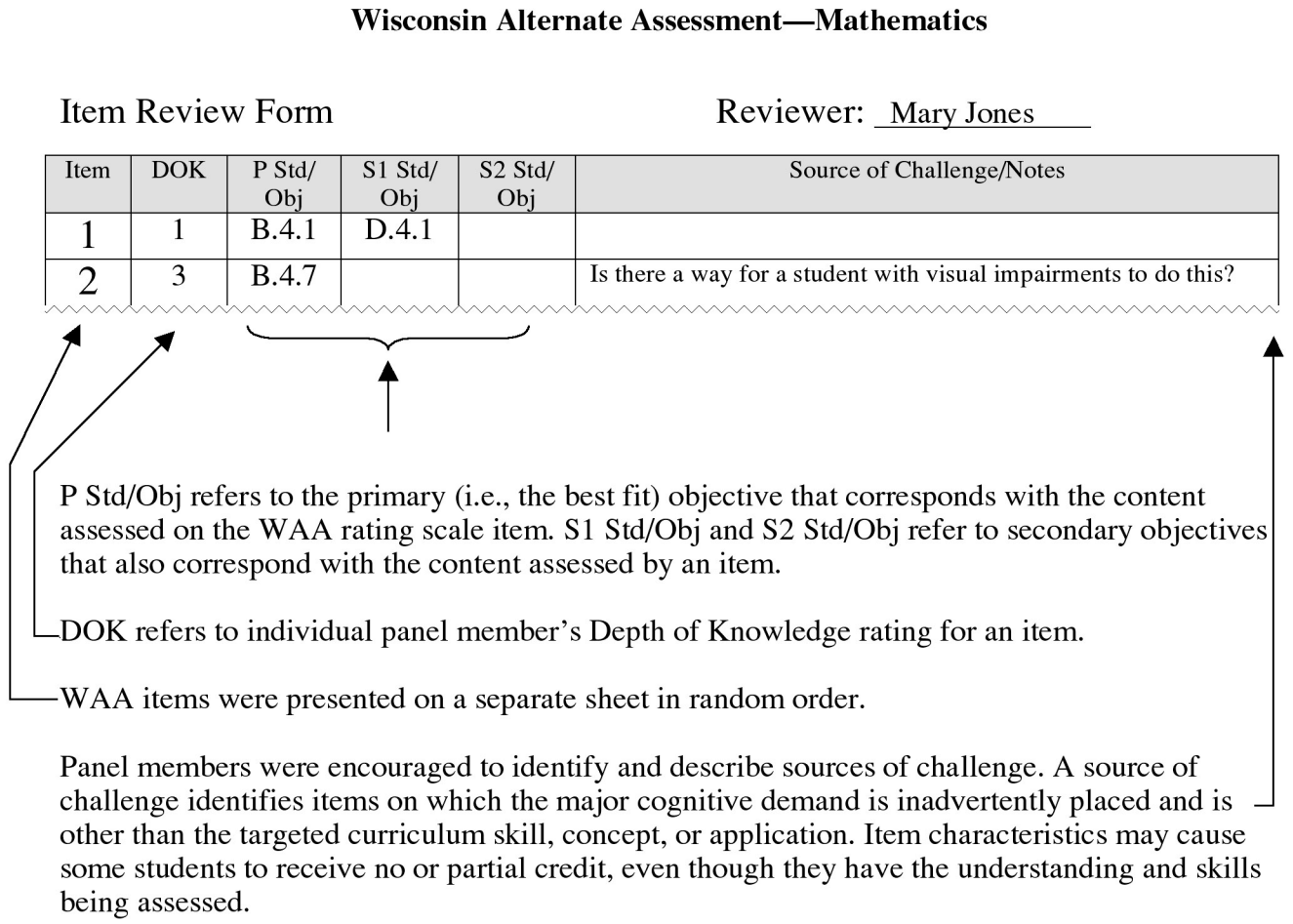

FIGURE 2. An example of a panel member's item review form.

respondence between these two documents. Subsequent analyses of the panel members' ratings resulted in descriptive statistics for the four criteria underlying Webb's alignment model: (a) categorical concurrence, (b) range-of-knowledge correspondence, balance of representation, and (d) depth-of-knowledge consistency. Webb's criteria for determining alignment between assessments and curricular expectations are outlined in Table 3.

Research Question and Statistical Analysis

The purpose of the WAA Alignment Institute was to determine whether the WAA adequately measures the skills and concepts represented in Wisconsin's Model Academic Standards. The alignment panel's responses were expected to indicate that the WAA generally conforms to Webb's model for alignment of assessments and curriculum expectations. Specifically, the expert panel was expected to indicate that each WAA subject domain scale meets the criteria for categorical concurrence, range of knowledge, and balance of representation but has a low overall depth-of-knowledge rating. Because the WAA is intended to be an assessment for students with significant disabilities, it was anticipated that the items would not demand the level of mastery expected from students who take the regular large-scale assessment. Specifically, WAA rating scale items were developed using the state's Alternate Performance Indicators (APIs), downward extensions of the state's Model Academic Standards intended to be more accessible for students with significant disabilities. Thus, the percentage of WAA items meeting the depth-of-knowledge criteria was expected to be less than 50%.

Results

Information about the extent to which the WAA adequately measures the skills and concepts represented in Wisconsin's Model Academic Standards was provided by an analysis of the following data gathered as part of the WAA Alignment Institute: (a) panel members' ratings of the depth-of-knowledge level of each objective in the academic standards, (b) panel

TABLE 3. Summary of Webb's Criteria for Alignment

Criteria	Description
Categorical concurrence	An assessment must have at least six items measuring content for each standard to demonstrate an acceptable categorical concurrence between the standard and the assessment. "The number of items, six, is based on estimating the number of items that could produce a reasonably reliable subscale for estimating students' mastery of content on that subscale. . . . Using a procedure developed by Subkoviak (1988) and assuming that the cutoff score is the mean and the reliability of one item is .1, it was estimated that six items would produce an agreement coefficient of at least .63" (Webb, 2002, p.4).
Range of knowledge	This criterion is based on the assumption that an assessment should test students' understanding or mastery of the majority of the knowledge (i.e., more than half the objectives) represented by any given standard (Webb, 2002). At least 50% of the objectives for a standard corresponded with at least one related WAA item on the basis of the ratings of Alignment Institute panel members.
Balance of representation	A balance index score was computed to judge the distribution of assessment items. "The balance index compares the proportion of items for each objective to proportion if the assessment items were evenly distributed among all possible objectives" (Webb, Horton, & O'Neal, 2002, p. 9). An index value of 0.7 or greater indicated that WAA items are distributed among all objectives to an acceptable degree without forming a monominal or binominal distribution of assessment items on objectives under a standard.
Depth of knowledge	"For consistency between the assessment and standard . . . at least 50% of the items corresponding to an objective had to be at or above the level of knowledge of the objective" (Webb, 2002, p. 4). Meeting this criterion suggests that a test demands the adequate depth of understanding and sufficient mastery of the knowledge and skills covered in the corresponding academic standards.

Note. Adapted from Webb (2002).

members' ratings of the depth-of-knowledge level of each item on the WAA rating scale; and (c) the objectives identified by panel members as corresponding with each WAA item.

Alignment Institute panel members reached consensus on the depth-of-knowledge level ratings for the objectives (i.e., performance standards) for the Reading, Language Arts, and Mathematics scales. Because of time constraints, the panels' most common depth-of-knowledge rating (i.e., the mode) was assigned to the objectives in social studies and science. This departure from the typical alignment process was necessary because of the scope of the alignment panel's work. In previous alignment investigations conducted by Webb and colleagues, panel members focused on one or two subject areas. In the current investigation, however, the alignment panel was completing depth-of-knowledge ratings for curricular objectives in five separate subject domains.

The decision to forgo the consensus process did not appear to affect significantly the reliability of panel members' ratings. In fact, across subject domains, panel members independently rated the depth-of-knowledge levels of individual WAA items with moderate to high consistency. The average measure of intraclass correlations (Shrout & Fliess, 1979), which compared the ratings of the 10 reviewers, was consistently 0.85 or higher (see Table 4).

One aspect of alignment between standards and assessments is whether both documents address similar content. The *categorical concurrence* criterion provides a very general analysis of the content match. Analysis of the results from the Alignment Institute indicates that the WAA scales demonstrated varying levels of categorical concurrence across subject domains (see Table 5). Academic standards with an acceptable level of categorical concurrence were judged by panel members to have at least six corresponding items on the WAA scale. The categorical concurrence of academic standards with five corresponding WAA items, according to panel members' ratings, was considered weak.

TABLE 4. Reliability of Alignment Panel's Depth-of-Knowledge Level Ratings of WAA Items

Subject domain	No. reviewers	No. items	Alpha	95% confidence interval
Reading	10	23	.95	.92–.98
Language arts	10	26	.94	.89–.97
Mathematics	10	29	.90	.83–.95
Science	10	21	.86	.74–.93
Social studies	10	29	.89	.82–.94

The WAA Language Arts and Science scales achieve categorical concurrence for less than 50% of academic standards. Although this result is less than optimal, it is important to emphasize that attaining the categorical concurrence criterion only indicates there are sufficient items to create subscales within a particular academic area. Because the WAA reports only total scale scores for each subject domain, meet-

TABLE 5. Categorical Concurrence for WAA Subject Domain Scales

Subject domain	Academic standards	Objectives (performance standards)	No. WAA items	No. Hits (*M*)	Percentage of academic standards acceptable
Reading	1	4	23	33.1	100
Language arts	5	14	26	39.3	40
Math	6	32	29	42.2	50
Science	8	41	21	26.2	13
Social studies	5	47	29	39.1	60

Note. Number of hits is the number of items panel members coded as corresponding to an academic standard. Categorical concurrence includes academic standards with weak categorical concurrence (i.e., mean number of hits is five to six).

ing this criterion was desirable but not necessary for determining the validity and usability of the assessment.

When standards and assessment are aligned, they cover a comparable breadth of knowledge. The *range-of-knowledge consistency* criterion measures the number of objectives (i.e., performance standards) that have at least one corresponding assessment item. This criterion is based on the assumption that an assessment should measure students' understanding or mastery of the majority of knowledge and skills (i.e., more than 50% of the corresponding objectives) represented by any given standard.

The results of the WAA alignment indicate that the range-of-knowledge criterion was met for the Reading and Language Arts scales (see Table 6). According to the panel members' ratings, 100% of the reading and language arts objectives (i.e., performance standards) had a corresponding WAA item. The mean number of objective hits (4.2 for reading and 1.1 for Language Arts Standard F) indicates that some panel members rated items as corresponding to the larger content standard without indicating a specific corresponding objective.

The range-of-knowledge criterion was also met for the Mathematics, Social Studies, and Science scales, although the panel members' ratings indicate that the WAA items only weakly met the criterion for the majority of standards. This result is attributable to the numerous academic standards for these subject domains and the relative brevity of the WAA subject domain scales. For example, the low levels of range-of-knowledge consistency between the Social Studies Standards B and E and the WAA Social Studies scale reflect the numerous objectives for those standards. Although the panel members' ratings indicated that multiple items on the WAA Social Studies scale corresponded to Standards B and E, the range of item hits was not expansive enough to strongly meet the range-of-knowledge criterion.

Whereas the range-of-knowledge criterion measures an assessment's breadth of content, the *balance of representation* is related to the degree of emphasis. Achieving balance of representation requires that assessment items be evenly spread among the objectives for a standard. If one objective is to be weighed more heavily on an assessment, teachers and policymakers need to be informed of this emphasis (Webb, 2002). The analysis of the balance of representation included the use of the *balance index* developed by Webb, which provides scores ranging from 0 (a large percentage of items correspond to only one or two objectives) to 1 (equal distribution of the items across objectives). The balance index compares the actual proportion of items for each objective to the proportion if assessment items were evenly distributed among all possible objectives (Webb et al., 2002). The balance of representation for all the subject domain scales was rated by the panel members as acceptable (see Table 7). This result is attributable to the concise format of the WAA rating scale in comparison to many individually administered standardized tests. The limited number of items for each subject domain scale demanded that the scale developers evenly distribute items among the objectives. The panel members' ratings confirmed that the item development process resulted in a well-balanced scale for assessing students' performance.

In addition to evaluating the correspondence between the skills and concepts addressed in the academic standards and on the WAA instrument, the Alignment Institute results also provided a measure of the complexity of knowledge required by both documents. *Depth-of-knowledge consistency* describes the alignment between the skills and understanding students are expected to possess as stated in the standards, and the skills and understanding necessary to complete the WAA successfully.

Although similar depth-of-knowledge ratings for curriculum objectives and assessment items are generally desirable, items on many alternate assessments may demand less depth of knowledge than items in the general education academic standards and on the corresponding large-scale assessment. WAA items represent the range of concepts and skills outlined in Wisconsin's Model Academic Standards, but these items are presented at a lower level of complexity that allows access for students with significant disabilities. Therefore, the WAA was not expected to demonstrate acceptable depth-of-knowledge consistency. The acceptance of a low overall depth-

TABLE 6. Range of Knowledge for WAA Subject Domain Scales

Subject domain	Academic standard	Objectives rated (*M*)	Objectives hit (*M*)	Objectives hit (%)	ROK acceptable?	Percentage of standards acceptable
Reading	A. Reading/Literature	4.2	4.2	100	Yes	100
Language arts	B. Writing	3	3.0	100	Yes	
	C. Oral language	3	3.0	100	Yes	
	D. Language	2	1.5	75	Yes	100
	E. Media & technology	5	2.7	54	Yes	
	F. Research & inquiry	1.1	1.1	100	Yes	
Math	A. Mathematical processes	5.2	4.5	87	Yes	
	B. Number operations/relationships	7.3	6.3	86	Yes	
	C. Geometry	4	1.7	43	Weak	66
	D. Measurement	5.1	4.4	86	Yes	
	E. Statistics/probability	5	1.1	22	No	
	F. Algebraic relationships	6.1	1.7	28	No	
Science	A. Science connections	5.3	2.3	44	Weak	75
	B. Nature of science	3.1	1.1	35	No	
	C. Science inquiry	8.2	4.8	59	Yes	
	D. Physical science	8.2	3.5	43	Weak	
	E. Earth & space science	8.1	3.3	41	Weak	
	F. Life & environmental science	4	2.3	58	Yes	
	G. Science applications	5.1	2.5	49	Weak	
	H. Science in social/personal perspectives	4.1	1.1	27	No	
Social studies	A. Geography	9.2	5.3	58	Yes	80
	B. History	10.1	2.4	24	No	
	C. Political science	6.2	4.2	79	Yes	
	D. Economics	7.1	4.8	67	Yes	
	E. Behavioral sciences	15.1	9.6	42	Weak	

Note. WAA = Wisconsin Alternate Assessment; ROK = range of knowledge. Objective rated (*M*) refers to the mean number of objectives for reviewers. If the number is greater than the actual number of objectives for a content standard, then at least one reviewer coded an item as corresponding to a content standard but was unable to find a specific objective that corresponded with the item. Objectives hit (*M*) refers to the mean number of objectives for which raters indicate at least one corresponding WAA item. Objective hit (%) refers to the percentage of the total number of objectives which had at least one item coded.

of-knowledge rating represents a departure from previous alignment studies using expert panel ratings (Webb, 2002; Webb et al., 2002). The results of the WAA Alignment Institute, however, indicate a generally acceptable level of depth-of-knowledge consistency for each subject domain scale (see Table 8).

Discussion

The purpose of this investigation was to provide evidence of the alignment between the Wisconsin Alternate Assessment (WAA) for students with disabilities and Wisconsin's Model Academic Standards. In addition, this investigation illustrates the application of a widely accepted alignment method for the development of an alternate assessment for students with significant disabilities. Although the results of the WAA Alignment Institute are promising, they represent only one component of the necessary validity evidence for the WAA rating scale as a measure of student achievement and performance on the skills and concepts represented in Wisconsin's Model Academic Standards.

Interpretation of Major Findings and Relation to Previous Research

The expert panel's responses during the WAA Alignment Institute indicate that the WAA rating scale is generally well aligned with the skills and knowledge represented by Wisconsin's Model Academic Standards. In fact, the performance of the WAA on the four criteria that constitute Webb's (1997) alignment model met or exceeded the performance of many

TABLE 7. Balance of Representation for WAA Subject Domain Scales

Subject domain	Academic standard	Objectives rated (*M*)	Balance index		Balance of representation acceptable?	Percentage of standards acceptable
			M	*SD*		
Reading	A. Reading/Literature	4.2	.79	.07	Yes	100
Language arts	B. Writing	3	.83	.07	Yes	
	C. Oral language	3	.88	.06	Yes	
	D. Language	2	.94	.11	Yes	100
	E. Media & technology	5	.86	.11	Yes	
	F. Research & inquiry	1.1	.98	.08	Yes	
Math	A. Mathematical processes	5.2	.73	.13	Yes	
	B. Number operations/relationships	7.3	.70	.05	Yes	
	C. Geometry	4	.96	.07	Yes	100
	D. Measurement	5.1	.84	.05	Yes	
	E. Statistics/probability	5	.98	.06	Yes	
	F. Algebraic relationships	6.1	.90	.11	Yes	
Science	A. Science connections	5.3	.95	.08	Yes	
	B. Nature of science	3.1	.98	.05	Yes	
	C. Science inquiry	8.2	.85	.06	Yes	
	D. Physical science	8.2	.92	.08	Yes	100
	E. Earth & space science	8.1	.85	.06	Yes	
	F. Life & environmental science	4	.98	.05	Yes	
	G. Science applications	5.1	.92	.08	Yes	
	H. Science in social/personal perspectives	4.1	1.00	.00	Yes	
Social studies	A. Geography	9.2	.84	.03	Yes	100
	B. History	10.1	.91	.10	Yes	
	C. Political science	6.2	.82	.10	Yes	
	D. Economics	7.1	.80	.05	Yes	
	E. Behavioral sciences	15.1	.82	.04	Yes	

Note. WAA = Wisconsin Alternate Assessment. Objective rated (*M*) refers to the mean number of objectives for reviewers. If the number is greater than the actual number of objectives for a content standard, then at least one reviewer coded an item as corresponding to a content standard, but was unable to find a specific objective that corresponded with the item. Balance index = $1 - (\sum | 1/(0) - I_{(k)} /(H) |)/2$; where 0 = Total number of objectives hit for the subject domain; $I_{(k)}$ = Number of items corresponding to objective (k); and H = Total number of items hit for the subject domain.

states' general education assessments using the same alignment method. In comparison, 60% of the special education experts surveyed as part of a recent analysis of alternate performance indicators in 42 states indicated that most states had not adequately assessed the general education curriculum standards (Browder et al., 2002).

The WAA rating scale was not expected to demonstrate acceptable depth-of-knowledge consistency using Webb's alignment procedures; in fact, meeting the depth-of-knowledge criterion could be considered an indication that some WAA items were too difficult for the population of students for whom the test was developed. The results of the WAA Alignment Institute, however, indicated a generally acceptable level of depth-of-knowledge consistency between the WAA and the majority of academic standards in reading, mathematics, and social studies. There are multiple plausible explanations for this unexpected result: (a) the wording of the WAA items is general enough to allow for more complex interpretations of the tasks; (b) panel members believed the items tapped the same skills and knowledge expected in the objectives in a way that made them accessible to students with severe disabilities; and (c) the skills and concepts expected in the state's academic standards may focus primarily on recall and simple application of knowledge.

Limitations of the Current Investigation and Directions for Future Research

Additional research is needed to understand the unexpected depth-of-knowledge consistency results for the WAA. Follow-up interviews might provide additional evidence about alignment panel members' perceptions of level of understanding

TABLE 8. Depth-of-Knowledge Consistency for the WAA Subject Domain Scales

Subject domain	Academic standard	Percentage of items below DOK for objectives	Percentage of items at (or equal to) DOK for objectives	Percentage of items above DOK for objectives	DOK acceptable?
Reading	A. Reading/Literature	42	47	11	Yes
Language arts	B. Writing	48	44	9	Yes
	C. Oral language	75	24	1	No
	D. Language	81	19	0	No
	E. Media & technology	41	52	7	Yes
	F. Research & inquiry	93	7	0	No
Math	A. Mathematical processes	45	53	2	Yes
	B. Number operations/relationships	6	89	5	Yes
	C. Geometry	36	59	5	Yes
	D. Measurement	2	88	10	Yes
	E. Statistics/probability	77	23	0	No
	F. Algebraic relationships	14	79	7	Yes
Science	A. Science connections	78	22	0	No
	B. Nature of science	70	25	5	No
	C. Science inquiry	33	47	20	Yes
	D. Physical science	43	53	5	Yes
	E. Earth & space science	24	64	11	Yes
	F. Life & environmental science	38	38	25	Yes
	G. Science applications	6	71	23	Yes
	H. Science in social/personal perspectives	27	64	9	Yes
Social studies	A. Geography	44	55	1	Yes
	B. History	25	56	6	Yes
	C. Political science	59	27	14	Weak
	D. Economics	82	17	1	No
	E. Behavioral sciences	59	34	6	Weak

Note. DOK = depth of knowledge.

and mastery required by both the WAA items and the state's Model Academic Standards. In addition, expert review of the Wisconsin's Model Academic Standards and WAA items by a panel of special education researchers and policymakers would provide additional evidence for the appropriateness and accessibility of these documents for students with significant disabilities.

Participants in the WAA Alignment Institute were primarily administrators from the Department of Public Instruction (DPI) Special Education Team and graduate students in educational psychology. Although the alignment panel had extensive understanding of testing and measurement, special education policy, and education accountability systems, the addition of other constituencies to the panel may have produced different alignment results. For example, the inclusion of special education researchers on an alignment panel may have provided additional insight into the instruction and assessment of students with significant disabilities.

Replication of the methods used in this investigation with other alternate assessments would provide additional evidence of the methods' applicability to the behavior ratings scales, checklists, and portfolios generally employed to assess the academic performance of students with significant disabilities.

Implications for Policy and Practice

Although the data examined in this investigation should be interpreted with caution, the findings can be used to guide efforts to refine and implement alternate assessment procedures. This investigation represents the initial application of a nationally recognized alignment procedure to an alternate assessment. The results suggest that Webb's alignment model can be meaningfully applied to alternate assessment, providing special education leaders and policymakers with a tool for gathering evidence of the validity of their state's assessment.

Based on the results of the WAA Alignment Institute, the investigators recommended that the WAA Leadership Team add items to the WAA Science scale to improve its alignment to the state's academic standards and to the IEPs and classroom curriculum of students with significant disabilities, which would improve the validity and utility of WAA results. Three new items were subsequently added to the WAA Science scale prior to the initial WAA implementation year (2002–2003). These items were developed from the state's APIs and were selected to represent content in areas of the science standards identified by the alignment study as not well represented on the WAA rating scale.

The results of this investigation also speak to the role of sequential development and expert review in promoting the alignment between the policy elements of curriculum, instruction, and assessment systems (Webb, 1997). Sequential development involves creating and accepting one policy element, which subsequently serves as a blueprint for additional policy elements. In the case of the WAA, sequential development was a central element in creating standards-based rating scale items. APIs, which were developed on the basis of the state's academic standards, served as the framework for the subsequent development of items for the WAA. Initial item development involved revising the APIs to include more objective behavioral descriptions and to enhance the likelihood that students' skills and knowledge could be demonstrated in a variety of ways. Expert review by experienced special educators and DPI leadership was then used to analyze the importance of each item as a learning outcome for students with significant disabilities. This group's resulting importance ratings guided the selection of standards-based rating scale items for inclusion on the WAA.

Conclusions

The current investigation provided content-related evidence for the validity of the WAA rating scale as a method of assessing the performance of students with significant disabilities. Specifically, the results suggest that performance on the WAA rating scale may constitute a valid index of achievement for the concepts and knowledge represented in Wisconsin's Model Academic Standards. Additional studies that provide evidence of the psychometric properties of the WAA rating scale will be necessary to establish the measure's validity.

In a presentation to the Alternate Assessment Forum at the CCSSO National Conference on Large-Scale Assessment, Ken Warlick (then with the Office of Special Education Programs, U.S. Department of Education) discussed federal provisions concerning students with disabilities and state and district assessments. In particular, Warlick (as cited in Quenemoen, Massanari, Thompson, & Thurlow, 200) affirmed the need to create alternate assessments that measure students' progress toward the goals and standards held for all students:

> The purpose of an alternate assessment should reasonably match, at a minimum, the purpose of the assessment for which it is an alternate. One might ask, "If an alternate assessment is based on totally different or alternate standards, or a totally separate curriculum, what is the alternate assessment an alternate to?" (p. 15)

This sentiment seems to be reflected in states' efforts to develop and refine their alternate assessment practices. "In 1999, 32% of states were using only functional skills for their alternate assessments with no link to state standards, by 2001 only 8% were doing so" (Browder, Fallin, Davis, & Karvonen, 2003, p. 259). As one respondent to a survey of national authorities on the education of students with significant disabilities commented, the "more these [alternate assessment] performance indicators are tied to [state] standards, the more important I believe they are" (Kleinert & Kearns, 1999, p. 105). In the same survey of experts, 20% of respondents questioned the Kentucky Alternate Assessment's focus on functional skill domains and recommended that these skills be integrated with participation in and mastery of general education curriculum standards. A 2003 National Center on Educational Outcomes (NCEO) survey of State Departments of Education indicated that most states were adhering to this advice, using academic content standards as the basis for their alternate assessments or linking functional skills to content standards (Thompson & Thurlow, 2001).

The results of the WAA Alignment Institute suggest that the WAA provides an alternate measure of the general subject domains included in Wisconsin's Model Academic Standards and on the WKCE, Wisconsin's general statewide assessment. Moreover, this investigation illustrates an approach to evaluating the alignment between alternate assessments and state content standards. Recent reviews of states' alternate assessment practices suggest that most states have not provided information on whether the skills included in their alternate assessments reflect the content of their state academic standards (Browder, Spooner, et al., 2003; Thompson & Thurlow, 2003). Thus, information derived from the alignment process described in this investigation could help demonstrate for policymakers, educators, and students' families that alternate assessments provide a meaningful alternate method of assessing students' progress on the skills and knowledge represented by state content standards (Thompson, Quenemoen, Thurlow, & Ysseldyke, et al., 2001).

Alignment to a state's general education academic standards, however, is only one aspect of creating a meaningful alternate assessment. Ysseldyke and Olsen (1997) suggested that, in addition to aligning to academic standards, alternate assessments should be relevant to curricula, measuring what students with significant disabilities are learning and doing in their classrooms. In many cases, the curriculum and instruction of students who participate in an alternate assessment differ significantly from those of other students. Therefore, test developers must determine the alignment between alternate assessments and the curriculum and instruction provided to

students with significant disabilities. The results of this investigation provide evidence for this relationship but also suggest the need for additional work to understand the correspondence between students' IEPs and the state's alternate assessment and academic standards. Strengthening this aspect of alignment will help ensure that students with significant disabilities are included in instructional improvement efforts and standards-based reform in a meaningful way.

REFERENCES

Browder, D. M., Fallin, K., Davis, S., & Karvonen, M. (2003). Consideration of what may influence student outcomes on alternate assessment. *Education and Training in Developmental Disabilities, 38,* 255–270.

Browder, D. M., Flowers, C., Ahlgrim-Delzill, L., Karvonen, M., Spooner, F., & Algozzine, R. (2002, April). *Curricular implications of alternate assessments.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Browder, D. M, Spooner, F., Algozzine, R., Ahlgrim-Delzell, L., Flowers, C., & Karvonen, M. (2003). What we know and need to know about alternate assessment. *Exceptional Children, 70,* 45–61.

Council of Chief State School Officers. (2002). *Models for alignment analysis and assistance to states.* Retrieved May 13, 2002, from http://www.ccsso.org/pdf/Alignment Models.pdf

Elliott, S. N. (2001). *Enhancing the Wisconsin Alternate Assessment project proposal.* Madison: Wisconsin Department of Public Instruction.

Elliott, S. N., Braden, J. B., & White, J. L. (2001). *Assessing one and all: Educational accountability for students with disabilities.* Arlington, VA: Council for Exceptional Children.

Ford, A., Davern, L, & Schnorr, R. (2001). Learners with significant disabilities: Curricular relevance in an era of standards-based reform. *Remedial and Special Education, 22,* 214–222.

Kleinert, H. L., & Kearns, J. F. (1999). A validation study of the performance indicators and learner outcomes of Kentucky's Alternate Assessment for students with significant disabilities. *The Journal of the Association for Persons with Severe Handicaps, 24,* 100–110.

Lane, S. (1999, October). *Validity evidence for assessments.* Paper presented at the Edward F. Reidy Interactive Lecture Series, The National Center for the Improvement of Educational Assessment, Providence, RI.

Quenemoen, R., Massanari, C., Thompson, S., & Thurlow, M. (2000). *Alternate assessment forum: Connecting into a whole.* Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved October 2, 2002, from http://education.umnedu/NCEO/OnlinePubs/Forum2000/ForumReport2000.htm

Shrout, P. E., & Fliess, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428.

Thompson, S. J., Quenemoen, R. F., Thurlow, M. L., & Ysseldyke, J. E. (2001). *Alternate assessment for students with disabilities.* Thousand Oaks, CA: Corwin Press.

Thompson, S., & Thurlow, M. (2003). *2003 state special education outcomes: Marching on.* Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved March 1, 2004, from http://education.umn.edu/NCEO/OnlinePubs/2003StateReport.htm

Thompson, S., & Thurlow, M. (2001). *2001 state special education outcomes: A report on state activities at the beginning of a new decade.* Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved December 15, 2003, from http://education.umn.edu/NCEO/OnlinePubs/2001StateReport.html

Thurlow, M., Olsen, K., Elliott, J., Yssledyke, J., Erickson, R., & Ahern, E. (1996). Alternate assessments for students with disabilities. *NCEO Policy Directions, 5,* 1–6.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (NISE Research Monograph No. 6). Madison: University of Wisconsin-Madison, National Institute for Science Education.

Webb, N. L. (2002, April). *An analysis of the alignment between mathematics standards and assessments for three states.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Webb, N. L., Horton, M., & O'Neal, S. (2002, April). *An analysis of the alignment between language arts standards and assessments in four states.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Ysseldyke, J. E., & Olsen, K. R. (1997). Putting alternate assessments into practice: What to measure and possible sources of data. *Exceptional Children, 65,* 175–186.